statistics

October 16, 2023

[1]:
```
# Loading libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
# Load dataset
df = pd.read_csv('weight-height.csv')
df.head()
```

[2]:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801

[3]:
```
# Library location
print(sns.__file__)
```

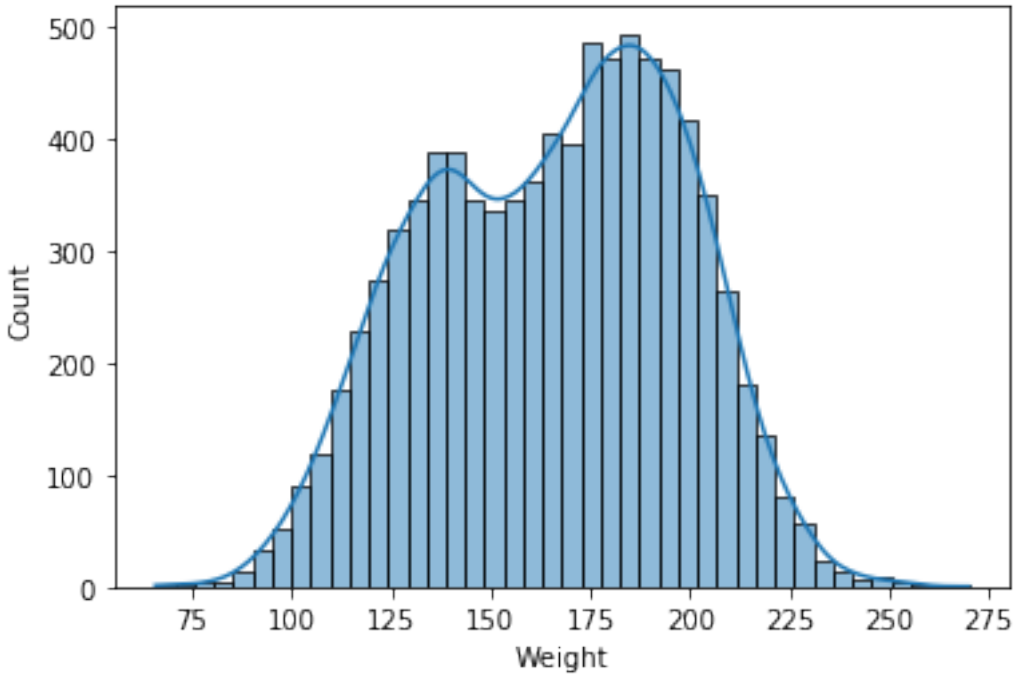
```
D:\anaconda3\lib\site-packages\seaborn\__init__.py
```

[4]:
```
# Seaborn Version
print(sns.__version__)
```

```
0.11.2
```

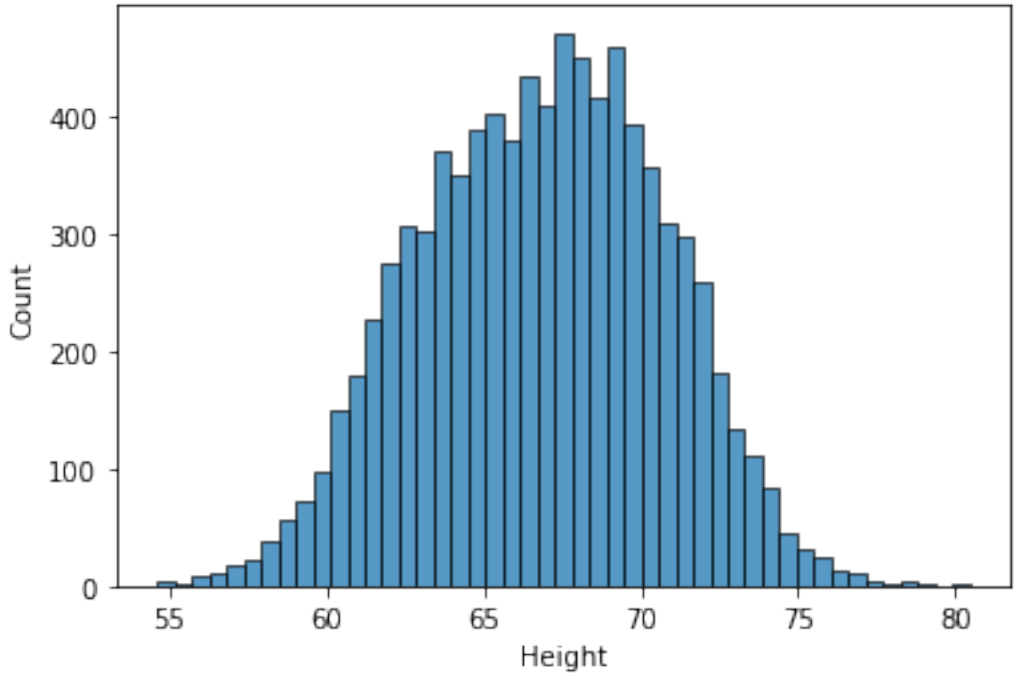
[5]:
```
sns.histplot(df.Weight,kde=True)          # Left skewed
```

[5]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x20464c54fc8>
```



[6]:
```
sns.histplot(df.Height)                    # Atmost normal Distributed
```

[6]: <matplotlib.axes._subplots.AxesSubplot at 0x20465ecec08>



0.1 Removing Outlier

[7]:
```
# concise summary
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8555 entries, 0 to 8554
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Gender  8555 non-null   object 
 1   Height  8555 non-null   float64
 2   Weight  8555 non-null   float64
dtypes: float64(2), object(1)
memory usage: 200.6+ KB
```

[8]:
```
# descriptive statistics
df.describe()
```

[8]:

	Height	Weight
count	8555.000000	8555.000000
mean	66.809925	165.632735
std	3.851454	32.043922
min	54.616858	65.780000
25%	63.957684	139.876803
50%	66.985923	168.521567
75%	69.604427	190.666305
max	80.450000	269.989698

[9]:
```
mean = df['Weight'].mean()
std = df['Weight'].std()

std_1 = mean + (1 * std)
std_2 = mean + (2 * std)
std_2
```

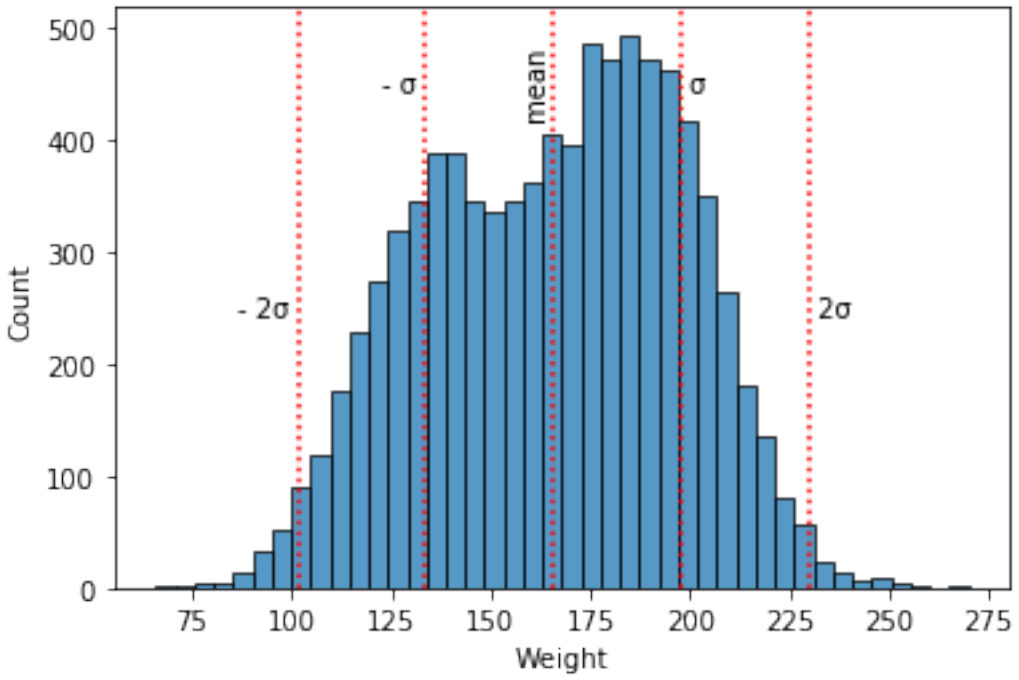
[9]: 229.72057976110685

[10]:
```
std_min_1 = mean - (1 * std)
std_min_2 = mean - (2 * std)
std_min_2
```

[10]: 101.54489089224487

[11]:
```
# Mean graph
plt.text(mean, 450, ' mean␣
 ↪',horizontalalignment='right',verticalalignment='center',rotation='vertical')
plt.axvline(x=mean, color='red', linestyle='dotted',label= 'mean')

# First Standard Deviation, ±
plt.text(std_min_1, 450, '- ␣
 ↪',horizontalalignment='right',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_min_1, color='red', linestyle='dotted',label= ' ')
plt.text(std_1, 450, ' ␣
 ↪',horizontalalignment='left',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_1, color='red', linestyle='dotted',label= '-  ')

# Second Standard Deviation, ±2
plt.text(std_min_2, 250, '- 2␣
 ↪',horizontalalignment='right',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_min_2, color='red', linestyle='dotted',label= '2 ')
plt.text(std_2, 250, ' 2␣
 ↪',horizontalalignment='left',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_2, color='red', linestyle='dotted',label= '- 2 ')

# plt.legend(loc = 'best')

sns.histplot(df.Weight)
```

[11]: <matplotlib.axes._subplots.AxesSubplot at 0x2046671e7c8>



[12]:
```
df[(df.Weight > std_min_2) & (df.Weight < std_2)]
```

[12]:

	Gender	Height	Weight
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
5	Male	67.253016	152.212156
…	…	…	…
8550	Female	60.483946	110.565497
8551	Female	63.423372	129.921671
8552	Female	65.584057	155.942671
8553	Female	67.429971	151.678405
8554	Female	60.921791	131.253738

[8345 rows x 3 columns]

[13]:
```
((8555-8551)/8555)*100
```

[13]:
```
0.0467562828755114
```

[26]:
```
mean = df['Height'].mean()
std = df['Height'].std()

std_1 = mean + (1 * std)
std_2 = mean + (2 * std)
std_2
```

[26]:
```
74.5128339922438
```

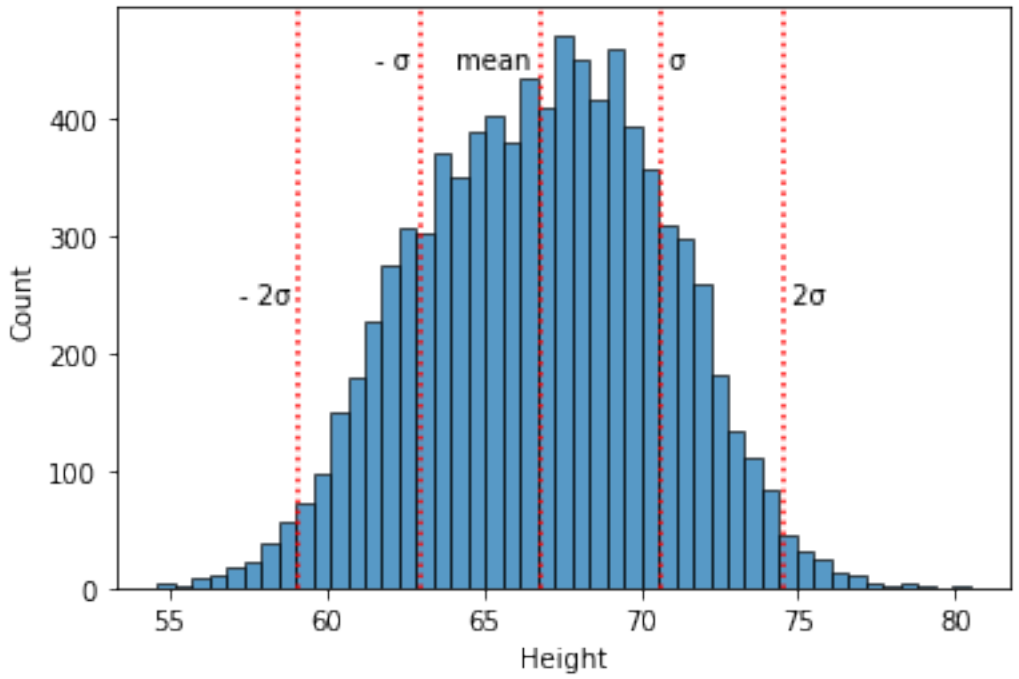
[27]:
```
std_min_1 = mean - (1 * std)
std_min_2 = mean - (2 * std)
std_min_2
```

[27]:
```
59.107016265934696
```

[29]:
```
# Mean graph
plt.text(mean, 450, ' mean␣
 ↪',horizontalalignment='right',verticalalignment='center',rotation='horizontal')
plt.axvline(x=mean, color='red', linestyle='dotted',label= 'mean')

# First Standard Deviation, ±
plt.text(std_min_1, 450, '- ␣
 ↪',horizontalalignment='right',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_min_1, color='red', linestyle='dotted',label= ' ')
```

```
plt.text(std_1, 450, ' ␣
 ↪',horizontalalignment='left',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_1, color='red', linestyle='dotted',label= '-  ')

# Second Standard Deviation, ±2
plt.text(std_min_2, 250, '- 2 ␣
 ↪',horizontalalignment='right',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_min_2, color='red', linestyle='dotted',label= '2 ')
plt.text(std_2, 250, ' 2 ␣
 ↪',horizontalalignment='left',verticalalignment='center',rotation='horizontal')
plt.axvline(x=std_2, color='red', linestyle='dotted',label= '- 2 ')

# plt.legend(loc = 'best')

sns.histplot(df.Height)
```

[29]: <matplotlib.axes._subplots.AxesSubplot at 0x2046693ce48>



[16]:
```
df[(df.Height > std_min_2) & (df.Height < std_2)]
```

[16]:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856

```
3       Male  71.730978  220.042470
4       Male  69.881796  206.349801
…        …        …          …
8550  Female  60.483946  110.565497
8551  Female  63.423372  129.921671
8552  Female  65.584057  155.942671
8553  Female  67.429971  151.678405
8554  Female  60.921791  131.253738

[8256 rows x 3 columns]
```

[17]:
```
((8555-8547)/8555)*100
```

[17]:
```
0.0935125657510228
```

[18]:
```
#removing outliers
df1 = df[(df.Height > std_min_2) & (df.Height < std_2)]
df1
```

[18]:
```
      Gender     Height      Weight
0       Male  73.847017  241.893563
1       Male  68.781904  162.310473
2       Male  74.110105  212.740856
3       Male  71.730978  220.042470
4       Male  69.881796  206.349801
…        …        …          …
8550  Female  60.483946  110.565497
8551  Female  63.423372  129.921671
8552  Female  65.584057  155.942671
8553  Female  67.429971  151.678405
8554  Female  60.921791  131.253738

[8256 rows x 3 columns]
```

[19]:
```
# Total removed rows
len(df)-len(df1)
```

[19]:
```
299
```

0.2 Z-Score

[21]:
```
df['z_scores'] = (df.Height - df['Height'].mean())/df['Height'].std()
df.head()
```

[21]:
```
  Gender     Height      Weight  z_scores
0   Male  73.847017  241.893563  1.827126
1   Male  68.781904  162.310473  0.512009
```

```
2   Male  74.110105  212.740856  1.895435
3   Male  71.730978  220.042470  1.277713
4   Male  69.881796  206.349801  0.797587
```

[22]:
```
df['z_scores'].max()
```

[22]:
```
3.541538687068086
```

[23]:
```
df['z_scores'].min()
```

[23]:
```
-3.165834495955818
```

[24]:
```
df2 = df[(df.z_scores > -2) & (df.z_scores < 2)]
df2
```

[24]:
```
      Gender     Height      Weight  z_scores
0       Male  73.847017  241.893563  1.827126
1       Male  68.781904  162.310473  0.512009
2       Male  74.110105  212.740856  1.895435
3       Male  71.730978  220.042470  1.277713
4       Male  69.881796  206.349801  0.797587
…        …         …          …         …
8550  Female  60.483946  110.565497 -1.642491
8551  Female  63.423372  129.921671 -0.879292
8552  Female  65.584057  155.942671 -0.318287
8553  Female  67.429971  151.678405  0.160990
8554  Female  60.921791  131.253738 -1.528808

[8256 rows x 4 columns]
```

[25]:
```
len(df) - len(df2)
```

[25]:
```
299
```